DOCUMENT RESUME

ED 133 347 TM 005 952

AUTHOR Huberty, Carl J.; Smith, Douglas U.
TITLE Variable Contribution in Discriminant Analysis.
PUB DATE [Apr 76]
NOTE 15p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Classification; *Discriminant Analysis; *Predictor Variables; *Statistical Analysis
IDENTIFIERS Criterion Variables

ABSTRACT

The purpose of this study was to determine which of six methods of ordering variables in a discriminant analysis yields subsets of variables that have the greatest discriminatory power. One method is based on univariate mean-square (or F) ratios, a second method on stepwise ordering, two methods on linear discriminant function (LDF) variable correlations, and two methods on standardized LDF coefficients. Real data on 80 graduate students in statistics were used. It was concluded that no single method was far superior to the others. Related findings are discussed, as are recommendations for subsequent research in this area. (Author/RC)

Variable Contribution in Discriminant Analysis

Carl J Huberty Douglas U. Smith

University of Georgia

Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976

Abstract

The purpose of this study was to determine which of six methods of ordering variables in a discriminant analysis yields subsets of variables that have the greatest discriminatory power. One method is based on univariate F's, a second method on stepwise ordering, two methods on LDF-variable correlations, and two methods on standarized LDF coefficients. Real data on 80 graduate students in statistics were used. It was concluded that no single method was far superior to the others. Related findings are discussed, as are recommendations for subsequent research in this area.

Variable Contribution in Discriminant Analysis

Introduction

Variables involved in a discriminant analysis may be considered criterion variables (in an "experimental" or group-separation problem), or predictor variables (in an ex post facto a group-classification problem). Thus, it would be helpful to be able to rank-order these variables in terms of their relative contribution to either group separation or to group classification accuracy. Such a rank ordering of variables would be informative for at least two reasons: (1) to aid in the interpretation of the discriminant analysis results for the data used, and (2) to discard variables for the purposes of subsequent research, thus lowering chances of misclassification given new data.

The problem of relative variable contribution has been studied from the one-group situation (Lutz, 1974), through the two-group situation (Cochran, 1964, Eisenbeis, Gilbert, and Avery, 1973), and to the more general k-group situation (Eisenbeis and Avery, 1972, Henschke & Chen, 1974; Huberty, 1975b). Some of these studies, and a few others, have explicitly attacked the related problem of variable selection (see Lachenbruch, 1975). The variable selection problem deals with determining a subset of the original set of variables of a given size the goal of which may be to select the subset that maximizes the difference between group mean vectors, or to select the subset that yields the greatest

classification accuracy. It is recognized that a subset determined by an index of relative contribution may not be the best subset in either of these two senses.

The focus of the present investigation was on the rank-ordering of variables with respect to the relative contribution made in classification accuracy. Six methods of ordering variables that have either been proposed or which have appeared in the literature were compared using real data. The purpose of the study, then, was to determine which of six methods is best, with "best" being defined in terms of the method which suggests subsets of the original set of variables having the greatest discriminatory power. As used in this study discriminatory power was assessed by the (internal/external) classification accuracy yielded by each subset.

Variable Ordering Methods

Two of the ordering methods selected for study are well known: (I) univariate mean-square (or F) ratios, and (II) (forward) stepwise discriminant analysis (BMD 7M in Dixon, 1973). Two other methods are intimately related to the eigenanalysis employed in deriving linear discriminant functions (LDFs). One of these (III) involves the correlations between each of the variables and each of the LDFs. For a given variable, the squares of these correlations are summed across the LDFs to obtain an index for that variable. These measures, the "communalities" for each variable, are only of interest when the number of variables is greater than one less than the number of groups (Cooley & Lohnes, 1971, p. 253). Another method (IV) involves the coefficients of each LDF that are applicable to standardized scores on the variables. For the ith variable a weighted composite of the

standardized coefficients (c_{ij}) is used; the jth weight is the eigenvalues (λ_j) associated with jth LDF: ($\sum_j \lambda_j c_{ij}$). The magnitude of this index is used to order the variables.

Finally, special cases of methods III and IV were considered in light of the data used in this investigation. Very often with more than three groups only one LDF is worthy of study. If so, method III simplifies to using (absolute values of) the leading LDF-variable correlation (Method V). And Method IV simplifies to using the standardized LDF coefficients (Method VI). Methods V and VI were considered to determine if the inclusion a "nonsignificant" LDF in using methods III and IV would substantially affect the discriminatory power of subsets of various sizes.

Data Analysis

The data used consisted of seven measures on 80 graduate students (Huberty & Smith, 1975). The seven measures were: age, two Graduate Record Examination (GRE) scores, two measures relating to undergraduate study in mathamatics/statistics, and two grade point averages. Group 1 ($n = 19$) consisted of those students who performed at the "A" level; group 2 ($n_2 = 37$) performed at the "B" level; and group 3 ($n_3 = 24$) performed at the "C" and below level. Discriptive data relative to the sample used in this investigation is given in Table 1.

Insert Table 1 about here

Initially, all seven variables were considered. Each of the ordering methods considered here (except for I, univariate Fs) call for the variables to be jointly normally distributed in the three populations, and for these populations to have a common covariance matrix. The constant

covariance structure was judged tenable since the value of Box's F statistic (Timm, 1975, p. 252) was less than unity. The value of Wilks's lambda was 0.460 which yielded $F = 4.81$ with $df = 14/142$, $p < .01$ $< .01$. The resulting eigenvalues were 1.077 and 0.046.

The rank-orderings of the variables according to all six methods are given in Table 2. As indicated by the resulting value of the

---

Insert Table 2 about here

---

coefficient of concordance ($W = 0.41$) there is moderate agreement among the orderings yielded by the six methods. Two pairs of rankings are of particular interest. First, it may be noted that consideration of the second LDF in using method III drastically modifies the ordering yielded by method V which considers only the leading LDF (rank-order correlation of -0.18). Second, it may be noted that the ordering for method IV is identical to that indicated by method VI. In light of the magnitude of the second eigenvalue (0.046), this is not too surprising. The standardized coefficients for the second LDF ranged from 0.064 to 2.609; the products of the second eigenvalue and these coefficients do not contribute a great deal (in a relative sense) to the composite, $\sum_j \lambda_j c_{i_j}$ -- the first product in the composite is the first eigenvalue (1.077) times coefficients ranging from 0.949 to 4.769 Thus five methods (I-IV) remained to be compared in terms of discriminatory power of suggested subsets of variables.

Subsets of variables at sizes 6, 5, 4, 3, 2, and 1 were specified according to each of the five methods. The discriminatory power of a subset of a given size based on each method was assessed using the results of a classification analysis. The classification statistic used

is one which provides posterior probabilities of group membership and which uses prior probabilities of group membership (15 in Huberty, 1975a). A linear classification rule was employed in this study since for each subset of variables considered, equality of covariance matrices was concluded.

Both internal and external classification results were obtained. The internal analysis being based on measures for those students on which basic statistics (mean vectors and covariance matrices) have been computed and then are resubstituted to obtain the values for the classification rules. In an external analysis statistics based on one set of students is used to classify "new" students. Even though a quadratic rule will yield greater internal accuracy when a linea. rule is considered appropriate, external classification based on a linear rule is often superior (see Huberty & Curry, 1975). The external analysis is essentially that suggested by Lachenbruch (1967).

Results

Proportions of correct classifications yielded by the internal and external analyses for the six subset sizes across the five ordering methods are given in Table 3. The rank-orderings of the classification

Insert Table 3 about here

proportions across the six subset sizes for the five methods produced moderate to low coefficients of concordance for the internal ($W = 0.43$) and external ($W = 0.05$) analyses. Thus, for external analyses, a considerable discordance of classification accuracy resulted. However, when examining the proportions two conclusions might be drawn: (1) Method IV (composite of weighted coefficients) yielded the highest proportion for all subset sizes, with the exception of subsets of size six. (2) For a

given subset size and across the five methods, the proportions do not differ greatly; for internal classification the maximum range of proportions was 0.075 (subset of size 2), and for external classification maximum range was 0.100 (subset of size 2).

An analysis of another data set (a three-group situation also) revealed that a second eigenvalue was also small relative to the first, and methods IV and VI yielded nearly identical variable rank-orderings --the only discrepency was that the ranks of the two poorest variables were interchanged. As was the case for results reported in the current paper, the results of an analysis using the second data set indicated that the consideration of a second (nonsignificant) LDF might be expected to modify the ordering yielded by method V which considers only the leading LDF.

There are some sidenotes of interest. First, proportions of correct internal classifications did not always increase with an increase in the number of variables entered into the analysis. Second, proportions based on an external analysis generally increased with a decrease in the number of variables entered analysis generally increased with a decrease in the number of variables entered, until the number decreased to one (Huberty & Curry, 1975). Third, once two variables were entered into the analysis the classification accuracy was not greatly affected, internally or externally, by the inclusion of more variables. This latter result may be a function of the size of the variable intercorrelations.

Discussion

Based on the results of this preliminary investigation, to infer that one of the six variable ranking methods is superior to the rest would be folly, indeed. There simply was not (that) much of a difference in the classification accuracy across the six methods. Essentially the

same general conclusion was reached when the second data set was analyzed (but not reported on here).

An additional real data situations need to be investigated with more group overlap, more criterion groups, different types of variables, and other variations, plus combinations of these variations. It may be difficult to locate real data sets having more than three groups and possessing some of the above variations for which the linear classification rule and most of the ordering methods proposed are appropriate. Hence, it may be desirable to conduct a Monte Carlo study, in which the true ordering of the variables is known, so as to determine which of the methods is best and which, if any, are good at all. Of equal, if not greater, interest is the variable ordering or selection problem when quadratic classification is appropriate (Lachenbruch, 1975).

REFERENCES

Cochran, W. G. On the performance of the linear discriminant function. Technometrics, 1964, 6, 179-190.

Cooley, W. W., & Lohnes, P. R. Multivariate data analysis. New York: Wiley, 1971.

Dixon, W. J. (Ed.). Biomedical computer programs. Berkeley, Calif.: University of California Press, 1973.

Eisenbeis, R. A., & Avery, R. B. Discriminant analysis and classification procedures. Lexington, Mass.: Heath, 1972.

Eisenbeis, R. A., Gilbert, G. G., & Avery, R. B. Investigating the relative importance of individual variables and variable subsets in discriminant analysis. Communications in Statistics, 1973, 2, 205-219.

Henschke, C. I., & Chen, M. M. Variable selection technique for classification problems. Educational and Psychological Measurement, 1974, 34, 11-18.

Huberty, C. J Discriminant analysis. Review of Educational Research, 1975, 45, 543-598. (a).

Huberty, C. J The stability of three indices of relative variable contribution in discriminant analysis. Journal of Experimental Education, 1975, 44, 59-64. (b)

Huberty, C. J, & Curry, A. R. Linear versus quadratic multivariate classification. Paper presented at the annual meeting of the American Educational Research Association, Washington, April 1975.

Huberty, C. J, & Smith, D. U. Measures of discrimination among achievement levels in statistics. Paper presented at the annual meeting of the American Educational Research Association, Washington, April 1975.

Lachenbruch, P. A. An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis. Biometrics, 1967, 23, 639-645.

Lachenbruch, P. A. Some unsolved problems in discriminant analysis. Institute of Statistics, University of North Carolina, Mimeo Series No. 1050, December, 1975.

Lutz, J. G. On the rejection of Hotelling's simple sample T^2. Educational and Psychological Measurement, 1974, 34, 19-23.

Timm, N. H. Multivariate analysis with applications in education and psychology. Belmont, Calif.: Brooks/Cole, 1975.

Table 1

Means, Standard Deviations[a], Univariate F's

and Within-Groups Correlation Coefficients

No.	Variable Name	Group 1 (n_1=19)	Group 2 n_2=37	Group 3 (n_3=24)	F	GREV	GREQ	UMSH	YCMS	UGPA	GGPA
1	Age	28.05 (4.74)	31.16 (7.43)	33.25 (7.13)	3.11	.25	-.09	.02	.79	-.09	-.06
2	GRE Verbal	558.21 (82.91)	505.35 (84.41)	467.92 (98.95)	5.51		.13	-.12	.25	.20	.15
3.	GRE Quantitative	626.84 (64.28)	543.95 (89.49)	474.50 (61.50)	21.08			.23	-.18	-.24	.02
4.	Undergraduate Mathematics/ Statistics Hours	21.84 (16.08)	15.35 (16.69)	10.88 (14.25)	2.54				-.28	-.12	-.11
5.	Number Years Since Last Mathematics/Statistics Course	6.16 (4.06)	10.22 (7.33)	12.04 (6.77)	4.44					-.15	.00
6.	Undergraduate GPA	3.33 (0.54)	2.99 (0.38)	2.81 (0.41)	7.66						.16
7.	Graduate GPA	3.75 (0.34)	3.72 (0.28)	3.51 (0.32)	4.58						

[a]Given in parentheses.

Table 2

Rank-Orderings of Variables

Method

	I (F's)	II (Stepwise)	III (Communalities)	IV (Weighted Coefficients)	V (r's)	VI (Coefficients)
Best	3	3	7	3	3	3
	6	6	3	6	6	6
	2	7	6	1	2	1
	7	4	5	4	5	4
	5	1	2	5	7	5
	1	2	1	2	1	2
Poorest	4	5	4	7	4	7

Table 3

Proportions of Correct Classifications[a]

No. Variables in Subset		Method I	II	III	IV	V	Maximum Difference
6	Internal	650	663	650	713	650	063
	External	613	588	613	588	613	025
5	Internal	650	675	650	657	650	025
	External	600	600	600	600	600	000
4	Internal	662	688	688	700	638	062
	External	600	600	625	638	613	038
3	Internal	675	688	700	675	675	025
	External	625	650	650	675	625	025
2	Internal	688	688	613	688	688	075
	External	675	675	575	675	675	100
1	Internal	513	513	488	513	513	025
	External	488	488	475	488	488	013

[a]Decimals are omitted.